Generative AI

Disrupting the Search Ecosystem

October 28, 2025

GenAI Disrupting the Search Ecosystem

- Thank you for your willingness to participate in our Annual Connections Event! Below you will find information for your session, including your personal join link. Should you have any technical difficulties the day of, I have provided my cell phone number below. Please call me with any questions or concerns on the day of your presentation.

AI Disruption

- AI services quickly replacing longstanding web search workflows
- Aggressive harvesting of content repository (including library catalogs and discovery services)
- Decreased human interactions
- Many needs are sufficed by AI summaries
- Real-time access increasingly occurs through interactive AI sessions (ChatGPT and others now have real time access to the web).

Disruption by generative AI

The rapid emergence of generative AI into the consumer and business sphere brings implications to libraries and their technology providers. Will AI short-circuit longstanding trajectories in the advancement of library management and discovery services? Library tech providers are making their initial forays into this sphere, but will these efforts bring substantial benefits to libraries and their user communities?

First phase of disruption

- AI has already disrupted the educational environment.
- Helping students and researchers generate text, images, and video for academic work
- Pressing ethical boundaries
- Libraries have integrated AI into instruction and academic support activities

Early phase of AI-based tools

- Chat services staffed by trained library workers will be extremely difficult to supersede
  - But why try unless the scale of demand is overwhelming?
- Deployment of AI-powered chatbots to replace or supplement traditional knowledge base and script-based services
- Early (largely unsuccessful) attempts to generate metadata for cataloging and description

Saf(er) AI: Retrieval augmented Generation

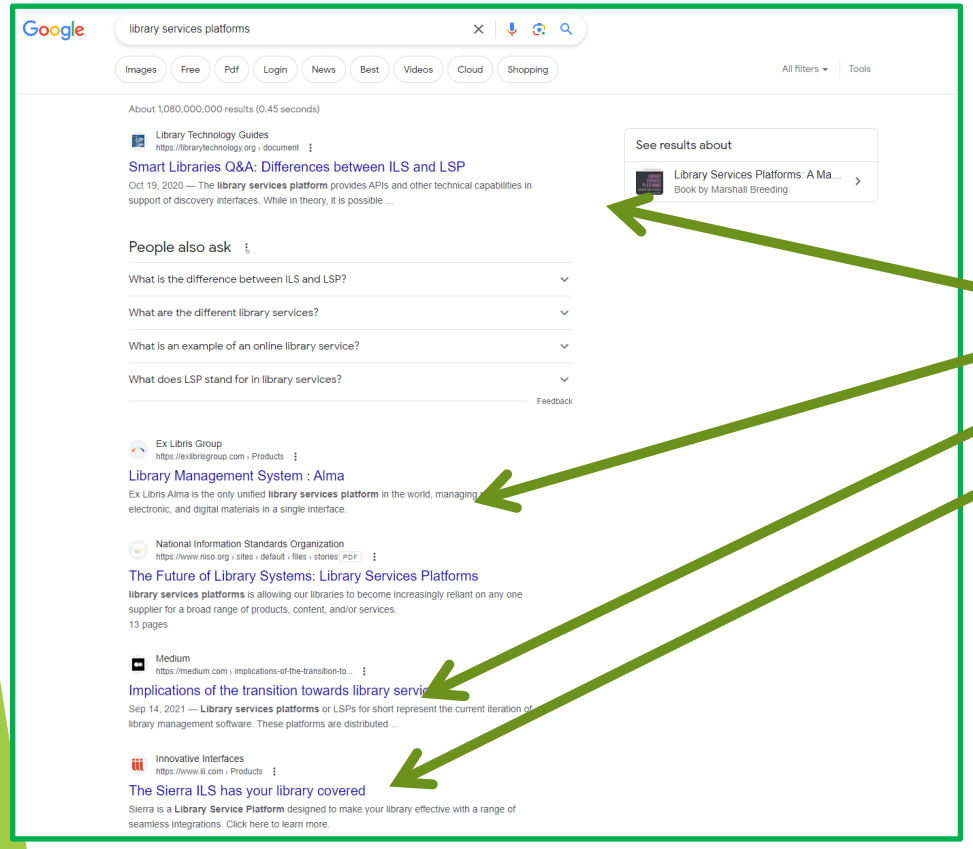
Google

library services platforms

Google Search presents delivers links to sponsored and organic results

Googlebot

More than 50% of requests are from search engine bots



PageRank Algorithm

Google Discovery Index

All Websites (except exclusions)

Classic Google Search presents organic blue links
(as well as sponsored links)

Google

library services platforms

Heavy overhead on web servers to enable Googlebot to populate discovery index

Google Generative Experience presents summaries but far fewer blue links

Googlebot

Google library services platforms

Converse | Images | Free | Pdf | News | Open source | Videos | Cloud | Shopping | All filters | Tools

Generative AI is experimental. Info quality may vary.

A library services platform (LSP) is a next-generation library management system that can manage both physical and electronic materials. LSPs are built on multi-tenant SaaS platforms that use cloud computing, web technologies, and discovery services. They support the work of library personnel and provide APIs and other technical capabilities to support discovery interfaces.

Here are some library services platforms:

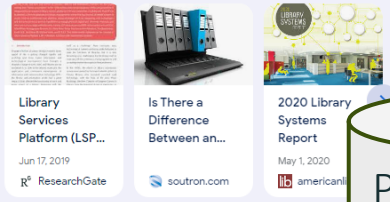
WorldShare Management Services (WMS)
Developed by OCLC, this LSP is for academic libraries. It provides a complete set of library management applications and platform services built on an open, cloud-based platform.

Alma
Developed by Ex Libris, this LSP is for academic, research, and national libraries.

EBSCO FOLIO
Developed by EBSCO, this LSP provides implementation, hosting, and support services.

Other library management software includes:

Schoolcanvas.com, Teachmint, Koha, Academics, Alexandria, LIBRARIAN.



Ask a follow up | What is the difference between ILS and LSP? | What are the examples of digital library service | What are

Library Technology Guides
https://librarytechnology.org › document

Smart Libraries Q&A: Differences between ILS and LSP

Oct 19, 2020 — The **library services platform** provides APIs and other technical capabilities in support of discovery interfaces. While in theory, it is possible ...

See results about

Library Services Platforms: A Ma...
Book by Marshall Breeding

PageRank Algorithm

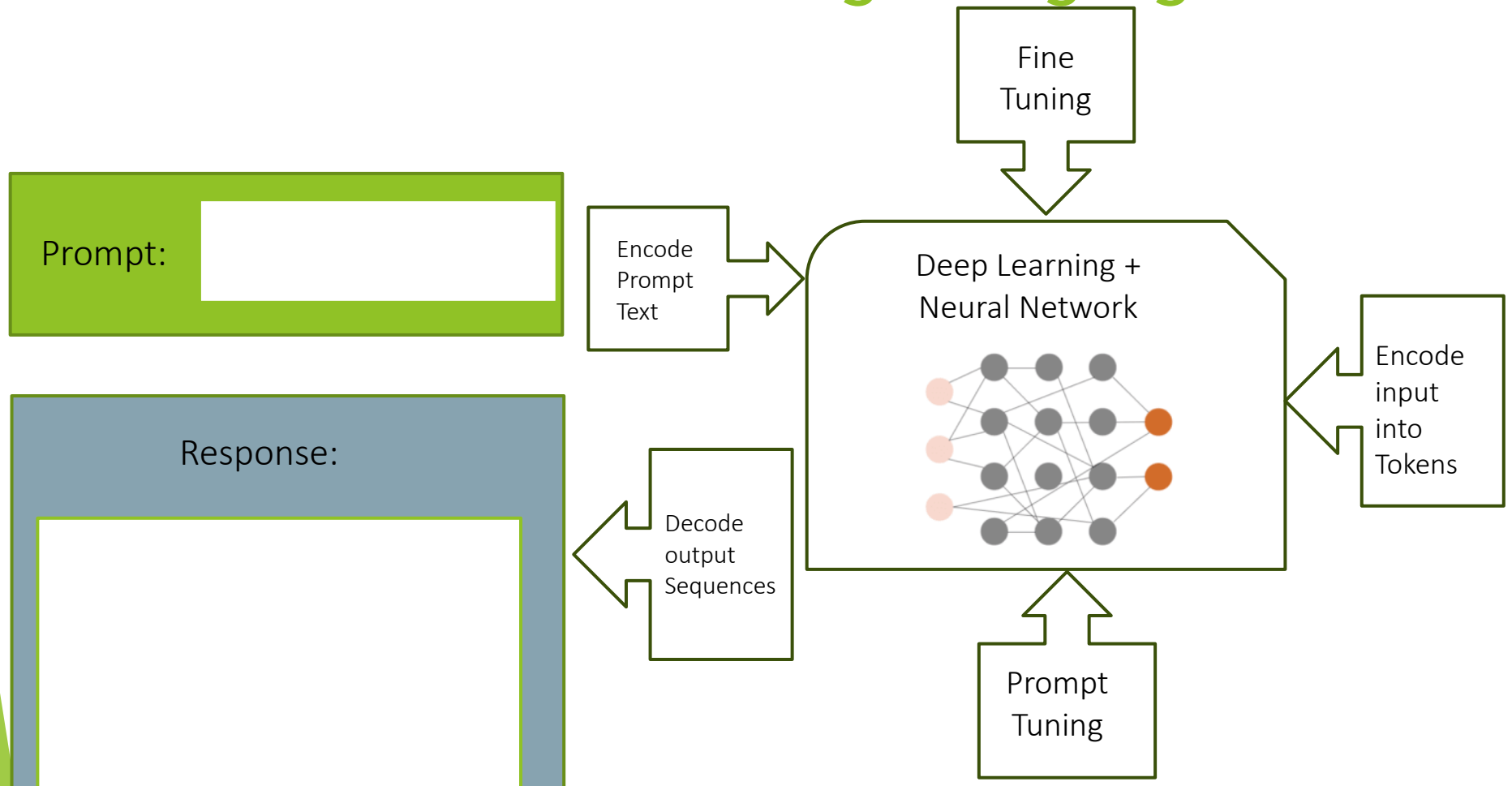
Google Discovery Index

All Websites (except exclusions)

?

Search Generative Experience

Generic Large Language Model



Over 250 billion pages spanning 16 years.
Free and open corpus since 2007.
Cited in over 10,000 research papers.
3–5 billion new pages added each month.
Primary training corpus in every LLM.
82% of raw tokens used to train GPT-3.
https://commoncrawl.org/

Retrieval Augmented Generation

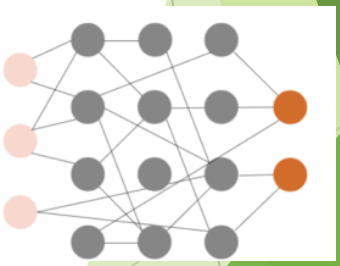
Prompt:

Response:

Vector Encoded Knowledgebase

Encode Prompt Text

Top Results



Text Generated by LLM

Content provided by reliable sources
Language generation provided by LLM

Avoids bias and errors by minimizing reliance on the LLM for facts and content

AI Powered Library Discovery

Beyond current index-based discovery services

Ongoing access to full text of scholarly articles based on keyword retrieval and relevancy ranking

New capabilities to return results based on concepts even when not found in the query text.

Cross-language searching

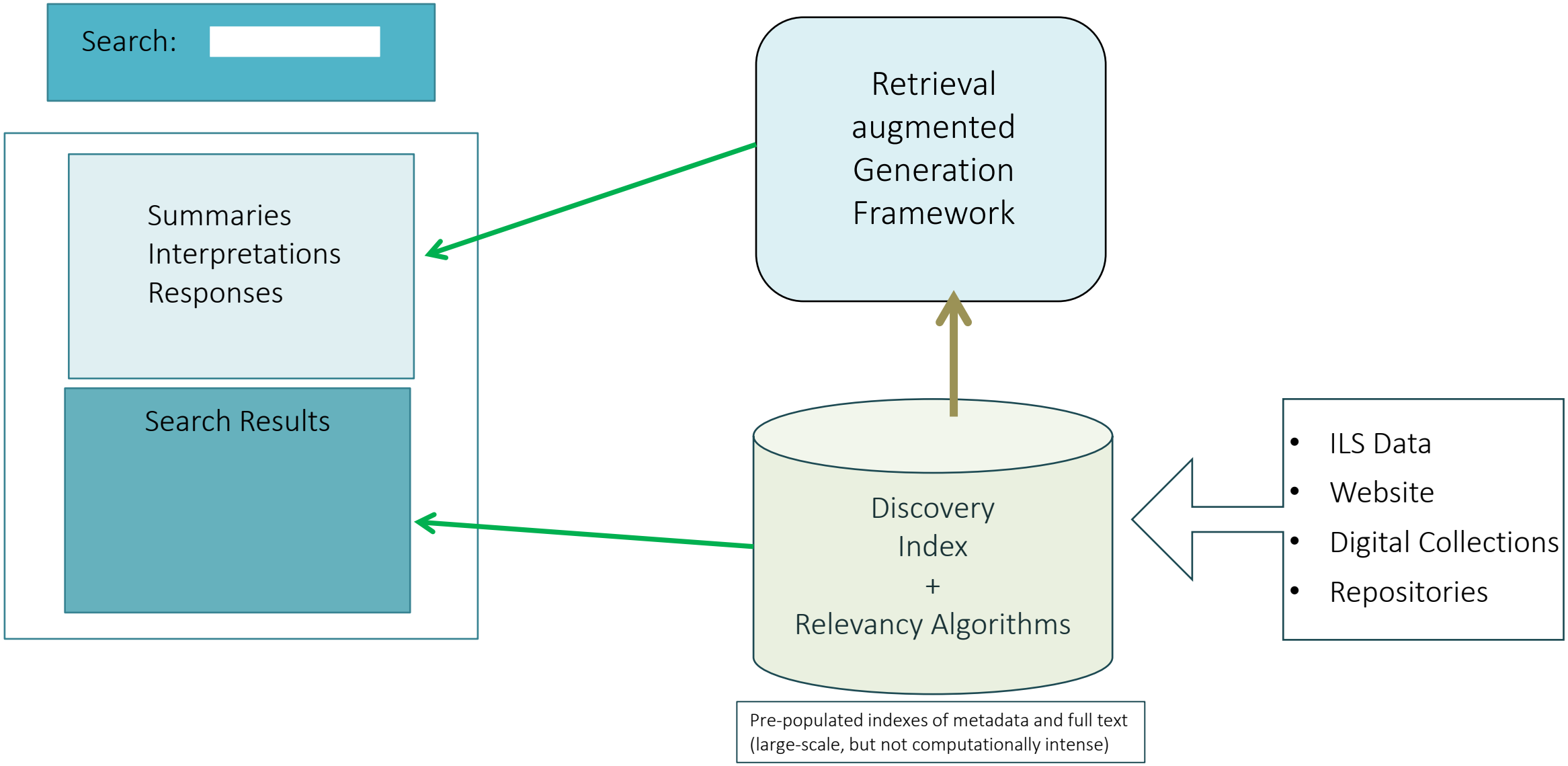
Summary and interpretation of results

Ability to explore topics with new tools and interfaces

Extract and cite relevant portions of text for incorporation in research papers

Many other capabilities that streamline user research and writing

Library Discovery Services



AI-powered Library Chat

Feasible to create conversational chat services based on native generative AI services geared to operational questions: hours, locations, policies, personnel.

Less feasible for chat services to respond to research questions that require live access to the web and online resources

RAG architecture can be used with Library Chat services for research questions if the framework includes access to adequate reference and research resources

Must be thoroughly tested to ensure high quality results

Use by library workers

Library patrons will increasingly be using AI. Librarians will need to understand these technologies to provide professional services

Use AI-based tools to improve productivity in selected areas of work

Incorporate in professional work and administrative tasks

Specific tasks that can be reviewed and revised before incorporating into official work products

The (dis)information Ecosystem

- An unfortunate reality:
  - Reliable information is difficult to access due to paywalls or other obstacles
  - Misinformation is abundant and easy to access
  - Significant impact on what is presented through search engines and incorporated into AI training data
  - Generative AI has the potential to create misinformation at scale
  - Impossible to distinguish text, images, and video created via AI
  - Distribution of misinformation at massive scale threatens intellectual freedom, scholarship, and democracy

Traffic patterns

- Increased harvesting by crawlers for AI services
- Increased activity for real-time AI interactions
- Search engine crawlers continue

Managing web traffic in the age of AI

- Create application firewall to filter incoming traffic
- Block mass harvesters
- Manage search engine bots
- Manage AI bots and chatbot agents

Identifying unwanted requests

- The http user-agent provides important clues regarding the validity of a page request. Although the user-agent can be forged to mis-represent the actual device making the request, patterns can be discerned that pragmatically inform whether the request is from an actual user or from an automated bot.
- This routine calls the DetectBrowser subroutine which returns the name and version number. The CheckUserAgent blocks requests from outdated browsers, from known web harvesting bots, and from unauthorized web crawlers.

Access

Library Technology Guides

Documents, Databases, News, and Commentary

Staff Home | Home | Libraries | Guides | Documents | Vendors | Products | News | Procurement

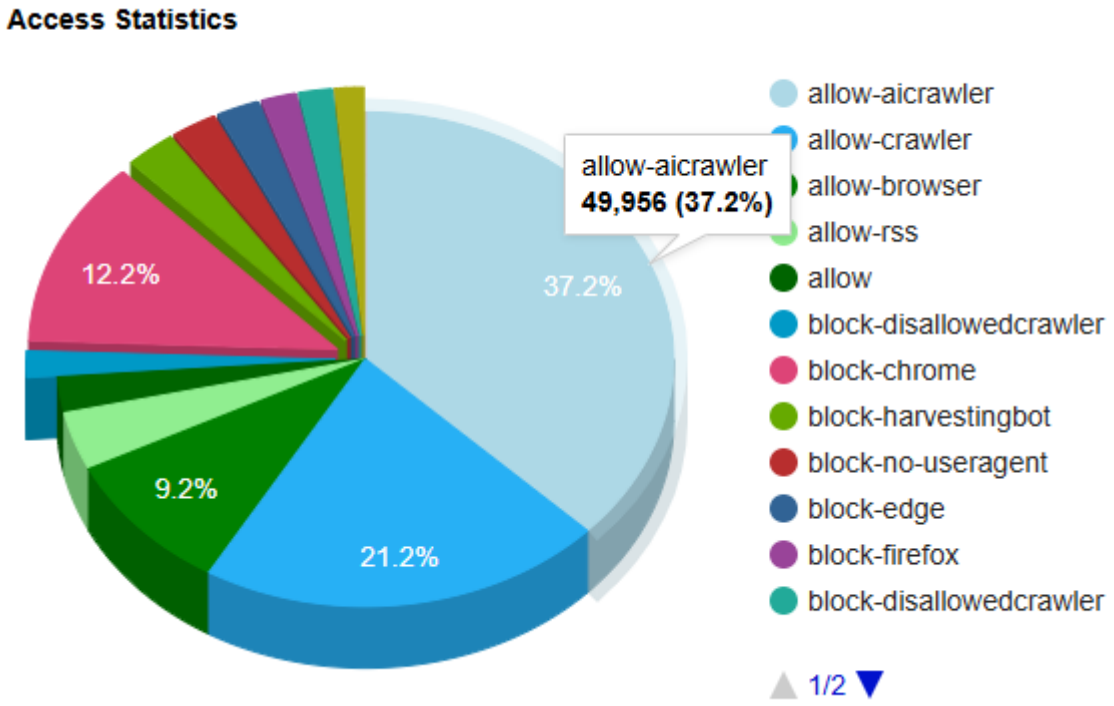
found **784** items where log entries in the last six minutes. Showing page **1** of **40**.

1 2 3 4 ...40 next >> search:

TimeStamp	IP	SubSystem	Page	Message	oper	cust	rec	count	Referrer
2025-10-24 14:44:45	20.169.6.230	lwc	Display Library		0	0	25645		OAISearchBot
2025-10-24 14:44:44	52.167.144.160	diglib	Blocked	Edge-112	0	0			
2025-10-24 14:44:43	52.167.144.160	lwc	Display List		0	0			BingBot
2025-10-24 14:44:43	172.182.224.14	lwc	Display Library		0	0	184758		OAISearchBot
2025-10-24 14:44:43	52.167.144.160	lwc	ProcessQuery	details	0	0		3	BingBot
2025-10-24 14:44:43	52.167.144.160	lwc	Display Library		0	0	21891		BingBot
2025-10-24 14:44:43	103.85.228.192	diglib	Blocked	Chrome-53	0	0			
2025-10-24 14:44:42	172.182.213.194	lwc	Display Library		0	0	17775		OAISearchBot
2025-10-24 14:44:42	202.40.194.41	rss	RSS News Feed		0	0			
2025-10-24 14:44:42	66.249.64.163	bib	displaytext		0	0	31098		GoogleBot
2025-10-24 14:44:41	20.14.99.109	lwc	Display Library		0	0	12638		OAISearchBot
2025-10-24 14:44:41	52.167.144.142	bib	displaytext		0	0	15750		BingBot
2025-10-24 14:44:40	52.167.144.142	lwc	Display Library		0	0	30190		BingBot
2025-10-24 14:44:40	52.167.144.142	lwc	Display Library		0	0	2125		BingBot
2025-10-24 14:44:40	20.169.6.229	lwc	Display Library		0	0	65392		OAISearchBot
2025-10-24 14:44:40	52.167.144.142	lwc	Display List		0	0			BingBot
2025-10-24 14:44:40	52.167.144.142	lwc	ProcessQuery	details	0	0		13	BingBot
2025-10-24 14:44:39	52.230.164.181	diglib	Index		0	0			GPTUserBot
2025-10-24 14:44:39	52.230.164.184	lwc	Display Library		0	0	199986		GPTUserBot
2025-10-24 14:44:39	20.14.99.109	lwc	Display Library		0	0	26479		OAISearchBot

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34

Traffic analysis



Top IP addresses		
IP address	Count	Domain
213.21.239.3	2585	m239-3.widemeshstaging.net
159.223.6.139	1824	cannot resolve
172.173.146.181	576	cannot resolve
109.239.229.214	525	cannot resolve
4.217.189.130	387	cannot resolve
74.235.234.40	386	cannot resolve
13.215.27.202	351	ec2-13-215-27-202.ap-southeast-1.compute.amazonaws.com
146.247.137.182	302	kepler3-g9.opoint.com
40.69.165.140	299	cannot resolve
52.176.0.226	241	cannot resolve
146.247.137.85	232	kepler7-g9.opoint.com
130.33.67.111	232	cannot resolve
52.138.216.201	196	cannot resolve
40.84.29.54	193	cannot resolve
13.213.239.22	189	ec2-13-213-239-22.ap-southeast-1.compute.amazonaws.com
146.247.137.125	177	newton21.opoint.com
18.136.146.78	164	ec2-18-136-146-78.ap-southeast-1.compute.amazonaws.com
74.176.63.5	116	cannot resolve
156.146.38.153	113	unn-156-146-38-153.cdn77.com
69.63.184.112	112	fwdproxy-ncg-112.fbsv.net
69.63.184.19	112	fwdproxy-ncg-019.fbsv.net

Ranked IP addresses from blocked requests

```
# Identify known bots. robots.txt controls inclusion or exclusion; The $Crawler flags requests as
our $Crawler = "";
    $Crawler = "GoogleBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /Googlebot/i);
    $Crawler = "GoogleOBot"            if ($ENV{'HTTP_USER_AGENT'} =~ /GoogleOther/i);
    $Crawler = "BingBot"               if ($ENV{'HTTP_USER_AGENT'} =~ /bingbot/);
    $Crawler = "InternetArchiveBot"    if ($ENV{'HTTP_USER_AGENT'} =~ /ia_archiver/i);
    $Crawler = "BaiduBot"              if ($ENV{'HTTP_USER_AGENT'} =~ /Baiduspider/);
    $Crawler = "AppleBot"              if ($ENV{'HTTP_USER_AGENT'} =~ /Applebot/i);
    $Crawler = "AmazonBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /Amazonbot/i);
    $Crawler = "YandexBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /YandexBot/i);
    $Crawler = "MediatoolkitBot"       if ($ENV{'HTTP_USER_AGENT'} =~ /Mediatoolkitbot/i);
    $Crawler = "NewsBlurBot"           if ($ENV{'HTTP_USER_AGENT'} =~ /NewsBlur/i);
    $Crawler = "SeekportBot"           if ($ENV{'HTTP_USER_AGENT'} =~ /SeekportBot/i);
    $Crawler = "AwarioBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /AwarioBot/i);
    $Crawler = "AwarioSmartBot"        if ($ENV{'HTTP_USER_AGENT'} =~ /AwarioSmartBot/i);
    $Crawler = "SeznamBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /SeznamBot/i);
    $Crawler = "intelx.ioBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /intelx\.io_bot/i);
    $Crawler = "DomCopBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /DomCopBot/i);
    $Crawler = "VelenPublicWebCrawler" if ($ENV{'HTTP_USER_AGENT'} =~ /VelenPublicWebCrawler/i);
    $Crawler = "generic-rb"            if ($ENV{'HTTP_USER_AGENT'} =~ /http\.rb/i);
    $Crawler = "SogouBot"              if ($ENV{'HTTP_USER_AGENT'} =~ /Sogou/i);
    $Crawler = "TrendictionBot"        if ($ENV{'HTTP_USER_AGENT'} =~ /trendictionbot/i);
    $Crawler = "ImagesiftBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /ImagesiftBot/i);
    $Crawler = "BarkrowlerBot"         if ($ENV{'HTTP_USER_AGENT'} =~ /Barkrowler/i);
    $Crawler = "PaqleBot"              if ($ENV{'HTTP_USER_AGENT'} =~ /Paqlebot/i);
    $Crawler = "CCBot"                 if ($ENV{'HTTP_USER_AGENT'} =~ /CCBot/i);
    $Crawler = "ECOsearchBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /ecoresearchCrawler/i);
    $Crawler = "ByteBot"               if ($ENV{'HTTP_USER_AGENT'} =~ /Bytespider/i);
    $Crawler = "ThreatViewBot"         if ($ENV{'HTTP_USER_AGENT'} =~ /ThreatView/i);
    $Crawler = "ArchiveBot"            if ($ENV{'HTTP_USER_AGENT'} =~ /ArchiveBot/i);
    $Crawler = "FlipboardBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /FlipboardProxy/i);
    $Crawler = "BuckBot"               if ($ENV{'HTTP_USER_AGENT'} =~ /Buck/i);

our $AICrawler = "";

    ## AI Bots
    $AICrawler = "GPTUserBot"            if ($ENV{'HTTP_USER_AGENT'} =~ /ChatGPT/i);
    $AICrawler = "OAISearchBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /OAI-SearchBot/i);
    $AICrawler = "GPTBot"                if ($ENV{'HTTP_USER_AGENT'} =~ /GPTBot/);
    $AICrawler = "PerplexityBot"         if ($ENV{'HTTP_USER_AGENT'} =~ /PerplexityBot/i);
    $AICrawler = "ClaudeBot"             if ($ENV{'HTTP_USER_AGENT'} =~ /ClaudeBot/i);
    $AICrawler = "ProtopageBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /Protopage/i);

our $BlockedCrawler = "";
    $BlockedCrawler = "TikTokBot"          if ($ENV{'HTTP_USER_AGENT'} =~ /TikTok/i);
    $BlockedCrawler = "YisouBot"           if ($ENV{'HTTP_USER_AGENT'} =~ /YisouSpider/i);
    $BlockedCrawler = "SEMrushBot"         if ($ENV{'HTTP_USER_AGENT'} =~ /SEMrushBot/i);
    $BlockedCrawler = "Qwantbot"           if ($ENV{'HTTP_USER_AGENT'} =~ /Qwantbot/i);
    $BlockedCrawler = "AliyunSecBot"       if ($ENV{'HTTP_USER_AGENT'} =~ /AliyunSecBot/i);
    $BlockedCrawler = "SemanticScholarBot" if ($ENV{'HTTP_USER_AGENT'} =~ /SemanticScholarBot/i);
    $BlockedCrawler = "ShopifyBot"         if ($ENV{'HTTP_USER_AGENT'} =~ /ShopifyBot/i);
    $BlockedCrawler = "ShapBot"            if ($ENV{'HTTP_USER_AGENT'} =~ /ShapBot/i);
    $BlockedCrawler = "Brightbot"          if ($ENV{'HTTP_USER_AGENT'} =~ /Brightbot/i);
    $BlockedCrawler = "ScreamingFrogBot"   if ($ENV{'HTTP_USER_AGENT'} =~ /Screaming/i);
    $BlockedCrawler = "DocsCrawlerBot"     if ($ENV{'HTTP_USER_AGENT'} =~ /DocsCrawler/i);
```

Access Strategy

IDENTIFY AND BLOCK MASS HARVESTERS

ALLOW SEARCH ENGINE BOTS

ALLOW SELECTED AI BOTS

CONTINUE TO EVALUATE BOT TRAFFIC

AI strategies for libraries

- Some libraries may want to employ AI technologies
  - AI-enhanced search
  - AI-powered chat services
  - AI workflow tools
- Even if the library does not use AI, it should have a strategy for engagement with AI ecosystem
  - Enable AI services to harvest data for training?
  - Enable AI services to interactively access library records and services?
- Disengaging with AI could isolate libraries from the information searching workflows that are becoming the dominant approach

Looking forward

- Difficult challenges persist in the political and business climate surrounding libraries
- Example: sudden collapse of Baker & Taylor
- Funding opportunities remain uncertain
- Polarized political situation often targets libraries and schools

- **Libraries must shape technology strategies to address prevailing realities**